Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 177 (2015) 169 – 177

Procedia Social and Behavioral Sciences

Global Conference on Contemporary Issues in Education, GLOBE-EDU 2014, 12-14 July 2014, Las Vegas, USA

Automatize Document Topic and Subtopic Detection with Support of a Corpus

Metin Turan[a*], Coskun Sönmez[b]

[a]*Computer Engineering Department,Yıldız Technical University, Esenler, 34220,Istanbul*
[b]*Computer Engineering Department,İstanbul Technical University,Maslak, 34469, Istanbul*

Abstract

In this article, we propose a new automatic topic and subtopic detection method from a document called paragraph extension. In paragraph extension, a document is considered as a set of paragraphs and a paragraph merging technique is used to merge similar consecutive paragraphs until no similar consecutive paragraphs left. Following this, similar word counts in merged paragraphs are summed up to construct subtopic scores by using a corpus which is designed so that we can find words related to a subtopic. The paragraph vectors are represented by subtopics instead of the words. The subtopic of a paragraph is the most frequent one in the paragraph vector. On the other hand, topic of the document is the most dispersive subtopic in the document. An experimental topic/subtopic corpus is constructed for sport and education topics. We also supported corpus by WordNet to obtain synonyms words. We evaluate the proposed method on a data set contains randomly selected 40 documents from the education and sport topics. The experiment results show that average of topic detection success ratio is about %83 and the subtopic detection is about %68.

Peer-review under responsibility of the Scientific Committee of GLOBE-EDU 2014.

Keywords: topic detection, text mining, document summarization

1. Introduction

The huge amount of documents needs to be processed, categorized and summarized to make life easier for researchers. Document processing age is still not capable of extracting information as a human reader. Moreover,

* Metin Turan. Tel.:+90 05327623554.
E-mail address: turanmetin@gmail.com

Peer-review under responsibility of the Scientific Committee of GLOBE-EDU 2014.
doi:10.1016/j.sbspro.2015.02.373

the importance of content in the document may also vary from one reader to another. In order to automatize the process needs to use both textual properties and resolve the language structure which is a really colicky operation with nowadays techniques.

Topic and subtopics detection is an important working area of the document processing. It gives us the relativeness of the document for a searched subject, analysis of the document structure (well-written or not), the boundary of the subtopics mentioned (for summarization) and the word relativeness in a hierarchical (events, subtopics and topics) order.

Different approaches of topic identification have been reported in the literature. The typical method in single document is text segmentation, which is to segment text based on the similarity of adjacent sentences and detect the boundary of subtopics (Choi, 2000; Hearst, 1997; Moens & Busser 2001; Zheng,Guo, Gong & Xue 2008). A popular method for topic identification in multiple documents is text clustering (Clifton, Cooley & Rennie, 2004; Radev, Jing, Stys & Tam, 2004).

A group of researchers worked on frequent word sequence (Liu, 2005; Yap, Loh, Shen & Liu, 2006) which is a sequence of words that appears in at least σ documents in a document set. It uses also frequent word pairs initially. Word clustering methods are used to find semantically related words. Words with similar distributions are grouped together (Pingoli, & Varma, 2007) and form a topic (Amini & Usunier 2007; Lin, Hovy, 2000). These models lack the semantic content information in documents and their summaries.

Concept-based classification (Blei, Ng, Jordan & Lafferty, 2011; Çelikyılmaz, Hakkani-Tur, 2011) is another approach applied to topic detection. Each sentence is associated with a multi-nominal distribution representing sparse-topic. A classifier separates the general and specific concepts (sLDA hidden topics). Finally, general topics sentences are selected for summary.

Probabilistic topic modeling is also a research area in this context (Gong, Qu & Tian, 2010). In this approach, each subtopic is associated with a set of document words and each sentence subtopics probability is evaluated. It is more than the classical TF-IDF and other textual properties calculation.

Agglomerative clustering algorithm is used to establish the hierarchical topic tree (Xu, Quan, Zhang & Wang, 2011). This approach divides documents into segments and then similar segments are merged up to a higher level. One of the old approaches applied for concept outline is Vector Space Model (VSM) (Ji, Lua, Wan & Gao, 2002). This is the fact that two words may have very different spelling but may be close semantically (“network”, “net”, “system”).

The latent semantic analysis (LSA) (Bellegarda, 2000; DeerWester, Dumais, Furnas, Landauer & Harshman, 1990) is proposed to extract the latent semantics from words and documents via singular value decomposition. Each document is projected to LSA space and represented by a low-dimensional vector where each entry acted as the weight on a specific topic (Hofmann, 1999).

A new hierarchical segmentation model (HSM) (Chien & Chueh, 2012) where the heterogeneous topic information in stream level and the word variations in document level are characterized. In this approach, the contextual topic information is incorporated in stream-level segmentation.

New event detection is an interesting research subject of news summarization. If a new includes novel information this should be checked as a recent event. In order to novelty detection, a cosine-similarity, a language-model and cover-coefficient methods are compared for the best result (Aksoy, Can, Kocberber, 2011).

Topic theme is a new approach proposed by Harabagiu and Lacatusu (2010) in order to evaluate the role of the correct topic detection on the summarization quality. They define a topic theme as any type of knowledge representation that encodes some facet of the knowledge pertaining to a topic. In their work, they compare the five known topic detection techniques with the new proposed two novel techniques.

The article is organized so that, the next chapter summarizes the inspired and related works. The chapter III includes research technique and details. The chapter IV informs about the data set and evaluation results. The last two chapters discuss the results and further work respectively.

2. Related Work

The first attempt to use word frequency as an indicator of document property is reached to the Luhn’s work (1958) which is accepted a milestone for document summarization. He suggested weighting sentences as a function

of high frequent words. After one decade later Edmundson (1969) suggested three methods (cue, title, location) to consider when evaluating the sentence weights.

TF is the best known and highly stable algorithm to calculate word frequencies in a document. It gives us what is the document mention about, or in another sense what is the most important issues in the document. Summarization is a task in order to reduce document size, finds out most important and non-redundant information and presents them in a readable rank. The most informative sentences include most frequent word/s. But, a question arises when a word frequency is not the most frequent one, but if it is seen most frequently through document units (sentence, paragraph). This is called Inverse Document Frequency (IDF) and accepted as an important indicator of document topic (Nakata, Ikeda, Ando & Okumura, 2002).

TFd_{iw} is count of how many times a word d_w appears in $document_i$ (frequency) while IDF measures what percentage of all documents in a collection contains that word. The IDF portion in formula *(1)*, N represents documents number in the document set, while Nd_W shows the number of documents contains word d_w

$$\mathrm{TFIDFd_{iw}} = \mathrm{TFd_{iw}} * \log\frac{\mathrm{N}}{\mathrm{Nd_w}} \quad (1)$$

The cut-off value for TF's is another interesting issue. What is the threshold value in order to select words which bring more important information.The informative (meaningful) words in the document are identified by some researchers (Balinsky H., Balinsky A. & Simske, 2011; Matsua, Ohsawa & Ishizuka, 2001).

To detect topics and event words, some researchers (M. Wang, C. Wang & Liu, 2008; M. Wang, X. Wang & Liu, 2009) tend to use TF-IDF values and dispersion values together. They believe if a word is an event word that word appears at one or several paragraphs of one document, but if it is topic it appears at whole documents. Formula *(2)* is dispersion value of a word w and denotes how frequently w appears across documents. m shows the number of documents in the document set, where $mean_w$ is the average of the word w in all documents.

$$\mathrm{DispD_w} = \sqrt{\frac{\sum_{i=1}^{m}(\mathrm{TFIDFd_{iw}} - \mathrm{mean_w})^2}{\mathrm{m}}} \quad (2)$$

Wang & Chan (2008) found similarities between paragraphs and choose the one with largest length as a component of the final abstraction result. If similarity values are higher than a threshold value, consecutive paragraphs can be merged or similar sentences can be extracted between these paragraphs. In Formula *(3)*, Sim (P1, P2) denotes the similarity between two paragraphs P1 and P2, V1 and V2 denote the vector used to represent P1 and P2.

$$\mathrm{Sim}(P_i, P_{i+1}) = \mathrm{Sim}(V_i, V_{i+1}) = \frac{V_i * V_{i+1}}{\sqrt{V_i * V_i} * \sqrt{V_{i+1} * V_{i+1}}} \quad (3)$$

Hierarchy corpus system can be sensible way to find topic which is called L-level PAM (Gong, 2010). It has four steps these are; documents set, super topic, sub-topic and word. Document set is the words in whole documents. Super topic is a general topic name like sports. Sub-topic contains super topics' specific words and word list involves sub topics specific words. For example; super topic is sports, subtopic contain sports branches like tennis and word list involves tennis ball, tennis racket and tennis court.

3. Work

3.1. Corpus Design

In order to memorize the keywords (topics, subtopics) for experimental subjects (education, sports), a corpus is constructed using a bulk of articles selected from Internet. It is a kind of a tree-structure where hierarchical relation between topic, subtopic and words are established. Fig. 1. shows the corpus design structure.

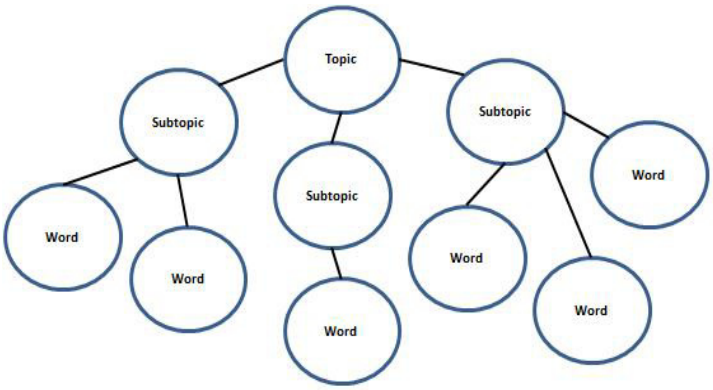


Fig. 1. Corpus Design

In corpus, 2 topics (education and sport), 37 subtopics (16 subtopics for education, 21 subtopics for sport) and total 2313 words are placed initially. Topics (education, sports) have broad range of subtopics and words to include in to the corpus all of them. These topics especially are preferred to test system success under insufficient corpus content.

In this project, we also get assistance from WordNet corpus to get word's synonyms in order to calculate term frequencies correctly. The assistance works like this; if a word is not contained by our corpus, we use WordNet synonym (synetsis) database to get synonyms of the word and match them from our corpus.

3.2. Topic and subtopic detection

The document is pre-processed initially. It includes:

- header selection
- sentence boundary identification
- stop-word elimination
- stemming
- paragraph boundary identification.

Special case for word frequency is that header words are calculated as twice. However it has no such effect when IDF is calculated (it is only used for topic detection).

Paragraph is accepted document unit and word vector representation is constructed only for paragraphs. Words in the paragraph vector these are related to the same subtopic (corpus is checked) are later huddled together to find subtopic counts for that paragraph. Finally, paragraph is defined as a list of subtopics (it is a new squeezed vector). The following example paragraph vector represents that paragraph P1 includes only $subtopic_c$ and $subtopic_e$.

$$\mathrm{V1}(0,0,\mathrm{subtopic_c},0\ \mathrm{subtopic_e},0)$$

Later each paragraph is processed by its consecutive paragraph for similarity. Cosine function is used to evaluate similarity. If there is similarity above a threshold value (determined 0.7) (M. Wang, X. Wang, 2009) then paragraphs are merged. A new paragraph vector is calculated for merged paragraphs. This operation goes on until there is no merging left.

The subtopics are obtained from the remained paragraphs. The most frequent subtopic in a paragraph vector is assumed the subtopic of that paragraph.

Paragraph Merging Example: Imagine that we have found four subtopics (X, Y, Z, W) in a document. In the case,

first paragraph P_1 contains 5 words for subtopic X and 3 words for Y subtopic, where second paragraph P_2 contains 2 words for subtopic X and 1 word for subtopic Z. The vector representations for P_1 *and* P_2 are given below.

$$V1\ (5_X, 3_Y, 0, 0) \qquad V2\ (2_X, 0, 1_Z, 0)$$

Similarity of these paragraphs using cosine function is calculated using the following formula *(3)*:

$$\mathrm{Sim}(P_1, P_2) = \frac{5*2+3*0+0*1+0*0}{\sqrt{5*5+3*3+0*0+0*0}*\sqrt{2*2+0*0+1*1+0*0}} = 0.76$$

Paragraph similarity is over 0.7 so these paragraphs are merged.

To calculate the dispersion value of all subtopics in the document found, we rearranged the formulas *(1)* and *(2)* given in the related work. In order to calculate TF and IDF values correctly only for one document, formulas are adopted to paragraph level. In formula *(4)*, P is the number of paragraphs in the document, where P_s shows the number of paragraphs contains subtopic s. i represents the ith paragraph in the document.

$$\mathrm{TFIDFp_{is}} = \mathrm{TFp_{is}} * \log\frac{P}{P_s} \qquad (4)$$

In order to calculate dispersion value of a candidate subtopic s, the dispersion formula given in *(2)* is also adopted to paragraph level. In formula *(5)*, m shows the number of paragraphs in the document, where $mean_s$ is the average of the subtopic s for all paragraphs in the document after subtopic detection.

$$DispP_s = \sqrt{\frac{\sum_{i=1}^{m}(TFIDFp_{is} - mean_s)^2}{m}} \qquad (5)$$

The topic of the document is obtained as the topic of the highest dispersive subtopic in the document.

4. Evaluation

4.1. Data set

A data set contains 40 documents from Internet web sites (www.elearningtech.blogspot.com, www.articlesbase.com, www.badminton-information.com) is randomly selected. Articles are about the sport and education experimental subjects. Sport documents topics include tennis, badminton, archery and general sport information. Education documents topics are science, preschool and e-learning. The documents selected have paragraphs between 6 and 9. The number of documents related to each subject is equal and 20.

A human reader extracted subtopics and topics of documents to compare with the output of the developed experimental system.

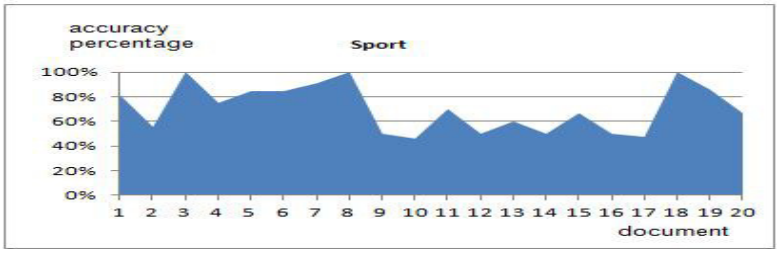
4.2. Evaluation

The theory is applied to an experimental system which is developed as a Computer Engineering Final Project (Turan, Kececi and Kesim, 2012). System has ability to get a document as an input, separate into paragraphs and finally gives the subtopics of left paragraphs and topic of that document.

System performance (correct subtopics and topics detection) is evaluated using two different working conditions: only with the Corpus, or Corpus plus WordNet.

The results obtained from experiments are presented as graphs in the following Fig. 2 and 3. In this graphs horizontal line shows the documents, where vertical line implies the accuracy percentage of subtopics detection in that document.

a



b

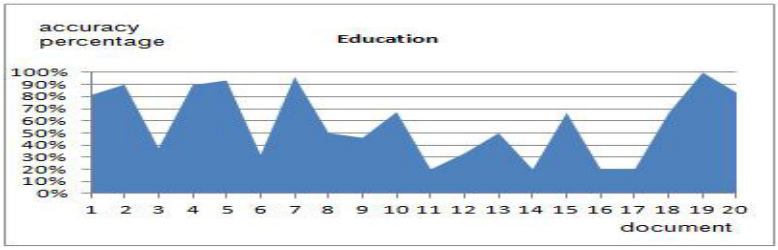
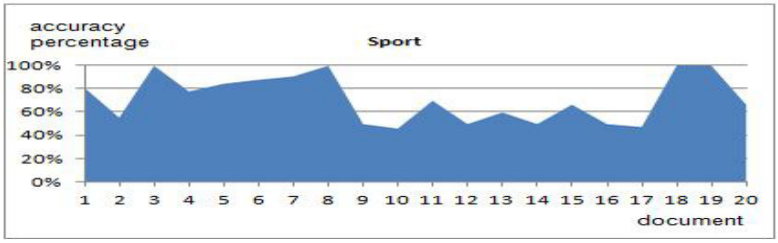


Fig. 2. Subtopic Detection only with the Corpus (a) Sport Dataset (b) Education Dataset

a



b

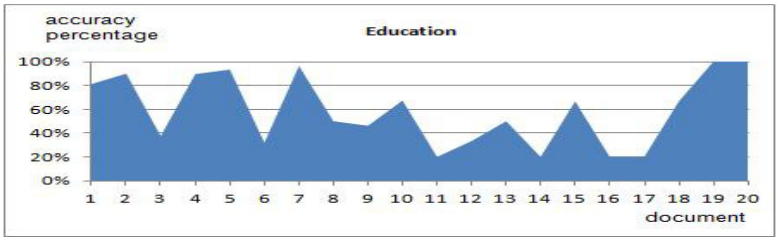
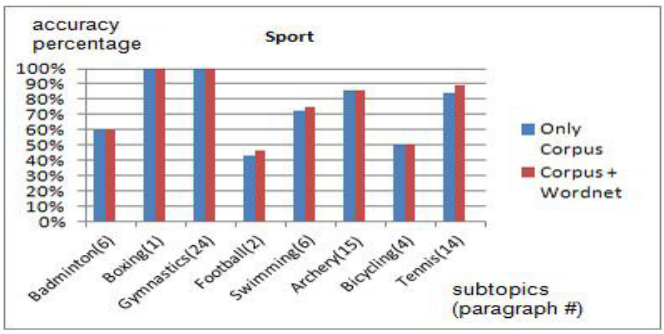


Fig. 3. Subtopic Detection Corpus plus WordNet (a) Sport Dataset (b) Education Dataset

Fig. 4. shows the details of the subtopics detection in Fig. 2. and 3. In this graphs horizontal line shows the subtopics and its total paragraph number in the data set, however vertical line outline the accuracy of these paragraphs subtopics detection. The results show that the average accuracy for subtopics detection in education subject is %61, however in sport it is %72.

a



b

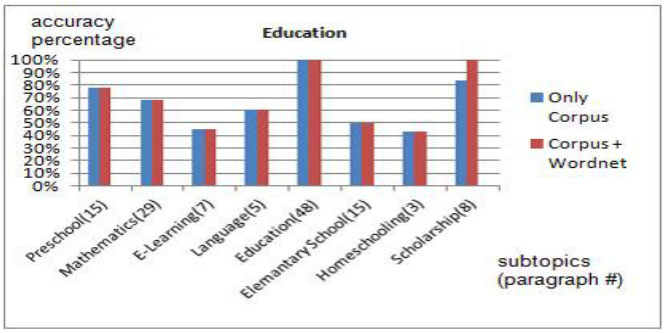


Fig. 4. Subtopic Details of Subjects (a) Sport Dataset (b) Education Dataset

a b

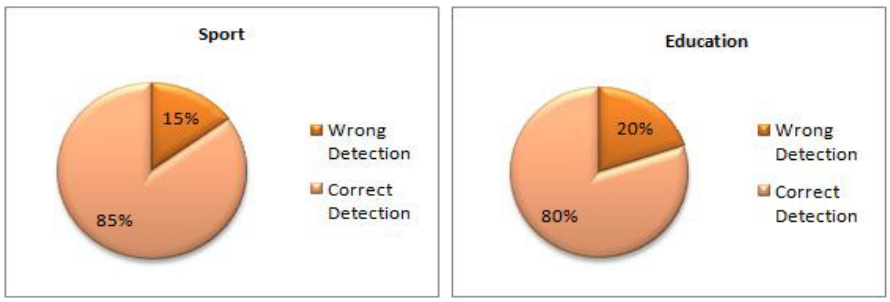


Fig. 5. Topic Detection (a) Sport Dataset (b) Education Dataset

The topics are determined for 40 documents in the data set. The Fig. 5. shows the results obtained. The results show that the average accuracy for topic detection is near %83.

Paragraph merging is another important evaluation point in that work. The number of merged consecutive paragraphs is calculated during documents processing and we got the number 156 over 271. The numbers show that merging paragraphs decrease the vector number significantly.

5. Conclusion

In this study, a corpus is used to find words these are related to same subtopic. By the way the word frequency is calculated using that strategy then the word list is decreased and a few subtopics is obtained for that document at the end of the process. It results a small sized vector to handle easily. Moreover, vector is constructed at the paragraph level, so the number of vectors is also limited to the paragraph number in that document at most. This also decreases the vector operations time.

A well-written document generally contains an order of subtopics. Consecutive paragraphs merging help us to group related text parts together. Experiments in this study have shown that nearly over %50 consecutive paragraphs are merged. Applying that technique, the number of paragraph vectors is decreased significantly. Briefly, paragraph merging is important for both subtopic detection and summarization.

WordNet usage affects our results insignificantly (Fig. 2. and 3.). The accuracy of the system depends on the organization and coverage of the corpus for used subject. The study is extended to extract words dynamically from the documents and select more valuable ones into the corpus.

Lots of research has proved that topic model can achieve more than 80% accuracy of finding the words correct topics in unsupervised ways (Li, Maccallum, 2006) Accuracy percentage of two other researches (Nakata, 2002; Wartena & Brusse, 2008) are %77 and %75 respectively. Comparison with the results in the literature, current system has accuracy over them and moreover it presents simple solution and produces speedy calculation.

6. Future work

One of the aims of that work is to automatize corpus generation. This is human-made corpus and can be extended to new subjects by the time. Such ability is added to system, but it still needs human assistance currently. In the future the corpus extension can be done heuristically.

The other research header is to summarize single or multiple documents with extracted information (topics and subtopics). Common subtopics in multiple documents can be ordered in importance to extract more valuable sentences in order to produce better extractive summary.

Observed that if a document is well-written it is so organized that same topic related paragraphs are placed next by next. In another words, the same topic paragraphs do not scattered in the document. In order to decide the document written style, that research can be extended using machine learning techniques.

References

Aksoy, C., Can, F. & Kocberber, S. (2011). Novelty detection for topic tracking, *Journal of the American Society for Information Science and Technology*, 777-795.

Amini, M.R. & Usunier, N. (2007). A contextual query expansion approach by term clustering for robust text summarization, *in Proc. of Document Understanding Conference*, 48-55.

Balinsky, H., Balinsky, A. & Simske, S. (2011). Document sentences as a small world, *in Proc. SMC*, 2583-2588.

Bellegarda, J. (2000). Exploiting latent semantic information in statistical language modeling, *in Proc. IEEE, 88*(8), 1279–1296.

Blei, D.M., Ng, A.Y., Jordan, M.I. & Lafferty, J. (2011). Latent dirichlet allocation, *Journal of Machine Learning Research*, 993-1022.

Celikyilmaz, A. & Hakkani-Tur, D. (2011). Concept-based classification for multi-document summarization, *Acoustics, Speech and Signal Processing (ICASSP)*, 5540-5543.

Chien, J.-T. & Chueh, C.-H. (2012). Topic-based hierarchical segmentation, *IEEE Transaction on Audio, Speech, and Language Processing*, *20*(1), 55-66.

Choi, F.Y.Y. (2000). Advances in domain independent linear text segmentation, *In Proceedings of the 1st North American Chapter of the Association for Computational Linquistics*, 26-33.

Clifton, C., Cooley, R. & Rennie, J. (2004). TopCat: data mining for topic identification in a text corpus, *IEEE Transactions on Knowledge and Data Engineering*, 949-964.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. & Harshman, R. (1990). Indexing by latent semantic analysis, *J. Amer. Soc. Inf. Sci.*, *41*(6), 391–407.

Edmundson, H.P. (1969). New methods in automatic extracting, *Journal of the ACM*, 264-285.

Gong, S., Qu, Y. & Tian, S. (2010). Subtopic-based multi-documents summarization, *Third International Joint Conference on Computational Science and Optimization*, 382-386.

Harabagio, S. & Lacatusu, F. (2010). Using topic themes for multi-document summarization, *ACM Transactions on Information Systems*, *28*(3), 1-46.

Hearst, M. A. (1997). TextTiling:segmenting text into multi-paragraph subtopic passages, *Computational Linguistics*, 33-64.

Hofmann, T. (1999). Probabilistic latent semantic indexing, *in Proc. ACM SIGIR*, 50–57.

Ji, H., Luo, Z., Wan, M. & Gao, X. (2002). Summarizing based on concept counting and hierarchy analysis, *IEEE SMC*.

Li, W. & Maccallum, A. (2006). Pachinko allocation: dag-structured mixture models of topic correlations, *In Proceeding of the 23rd International Conference on Machine Learning*, 577-584.

Lin, Y.-C. & Hovy, E.H. (2000). The automated acquisition of topic signatures for text summarization, *In COLING*, 495-501.

Liu, Y. (2005). *A Concept-based Text Classification System for Manufacturing Information Retrieval*. (PhD thesis). National University of Singapore, Singapore.

Luhn, H.P. (1958). The automatic creation of literature abstracts, *IBM Journal of Research and Development*, 159-165.

Matsuo, Y., Ohsawa, Y. & Ishizuka, M. (2001). A document as a small world, *Lecture Notes in Computer Science*, 444-448.

Moens, M.-F. & Busser, R. D. (2001). Generic topic segmentation of document texts, *In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 418-419.

Nakata, T., Ikeda, T., Ando, S. & Okumura, A. (2002). Topic detection based on dialogue history, *COLING '02 Proceedings of the 19th International Conference on Computational Linguistics*, 1-7.

Pingali, P., K, R. & Varma, V. (2007). IIIT Hyderabad at DUC 2007, *In Proc. of Document Understanding Conference*, NIST.

Radev, D.R., Jing, H., Stys, M. & Tam, D. (2004). Centroid-based summarization of multiple documents, *Information Processing and Management*, 193-207.

Turan, M., Kececi, O. & Kesim, A.E. (2012). *Article (document) Topic and Subtopic Detection*. (under-graduate thesis). İstanbul Kültür University, İstanbul.

Wang, H.-C. & Chan, Y.-C. (2008). On the abstraction and presentation of multi-source knowledge, *Proceedings of the Seventh International Conference on Machine Learning and Cybernetics*, 3307-3309.

Wang, M., Wang, C. & Li, Z.Z. (2008). Multi-document summarization based on word feature mining, *International Conference on Computer Science and Software Engineering*, 743-746.

Wang, M., Wang, X. & Li, C. (2009). Extracting multi-document summarization based on local topics, *Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, 238-241.

Wartena, C. & Brusse, R. (2008). Topic detection by clustering keywords, *DEXA '08 Proceedings of the 2008 19th International Conference on Database and Expert Systems Application*, 54-58.

Xu, Y.-D., Quan, G.-R., Zhang, T.B. & Wang, Y.-D. (2011). Used hierarchical topic to generate multi-document automatic summarization, *Fourth International Conference on Intelligent Computation Technology and Automation*, 295-298.

Yap, I., Loh, H.T., Shen, L. & Liu, Y. (2006). Topic detection using MFSs, *Proceedings of the 19th International Conference on Industrial&Engineering Applications of Artificial Intelligence & Expert Systems (IEA/AIE), Lecture Notes in Computer Science*, 342-352.

Zhang, Y. , Guo, J., Gong, H. & Xue, Z. (2008). Single-document automatic abstracting system based on topic partition, *International Symposium on Intelligent Information Technology Application Workshops*, 280-283.